

Halloween Is Coming: Let Enterprise Search Find the Ghouls in Your Data

Ghouls on Halloween Night trick-or-treating for candy? Good. Ghouls popping up without warning in enterprise data in the all-important 4th Quarter of the year? Bad. But enterprise search can stay one step ahead of these data ghouls.

So what are these data ghouls?

I view a data ghoul as anything buried in enterprise data that can re-emerge unexpectedly causing trouble. But enterprise search can help to proactively find such data ghouls before they do too much damage. dtSearch® for example can run in a classic Windows client-server environment, in an “on premises” web server capacity, or from the cloud such as Azure or AWS to let multiple end-users instantly and concurrently search terabytes of data.

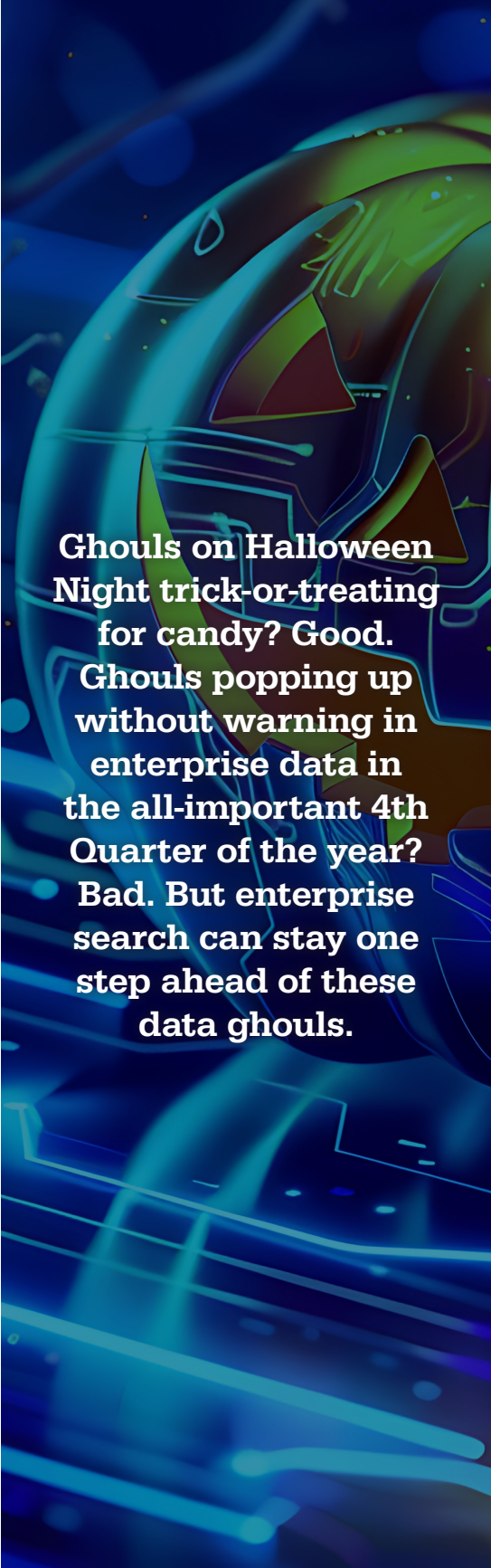
How does enterprise search work?

Indexing is a prerequisite to instant concurrent searching. However, indexing couldn't be easier. Just point to the folders, email archives, etc. you want the indexer to cover, and the software will take it from there. The folders can consist of local files or remote OneDrive Office 365 files or SharePoint attachments that, while appearing in the Windows folder system, are cloud-based. Whatever the file configuration, how enterprise search parses each item is critical for locating hidden data ghouls.

How does parsing work?

To parse each file, enterprise search goes directly to the binary version, whether local or cloud-based. Every file type—PDF, Microsoft Word, Access, Excel, PowerPoint,

Article contributed
by dtSearch®



**Ghouls on Halloween
Night trick-or-treating
for candy? Good.
Ghouls popping up
without warning in
enterprise data in
the all-important 4th
Quarter of the year?
Bad. But enterprise
search can stay one
step ahead of these
data ghouls.**

OneNote, Outlook, Exchange, etc.—has its own data specifications that the search engine needs to take into account for accurate parsing. dtSearch relies on the binary format to determine each item's file type. That way, an Access database that someone saves with a .DLL extension or a PDF that someone saves with a .ONE extension will not affect the indexer's ability to correctly parse the item.

Are there other benefits to parsing the binary formats of files?

The binary format also lets the indexer plow through recursively nested files like an email with a ZIP or RAR attachment with an Excel spreadsheet that itself includes a Word document. When viewing a file in its native application, you might click and click and still never run across certain obscure metadata. But all metadata is fully apparent in the binary format and thus easily accessible to the indexer.

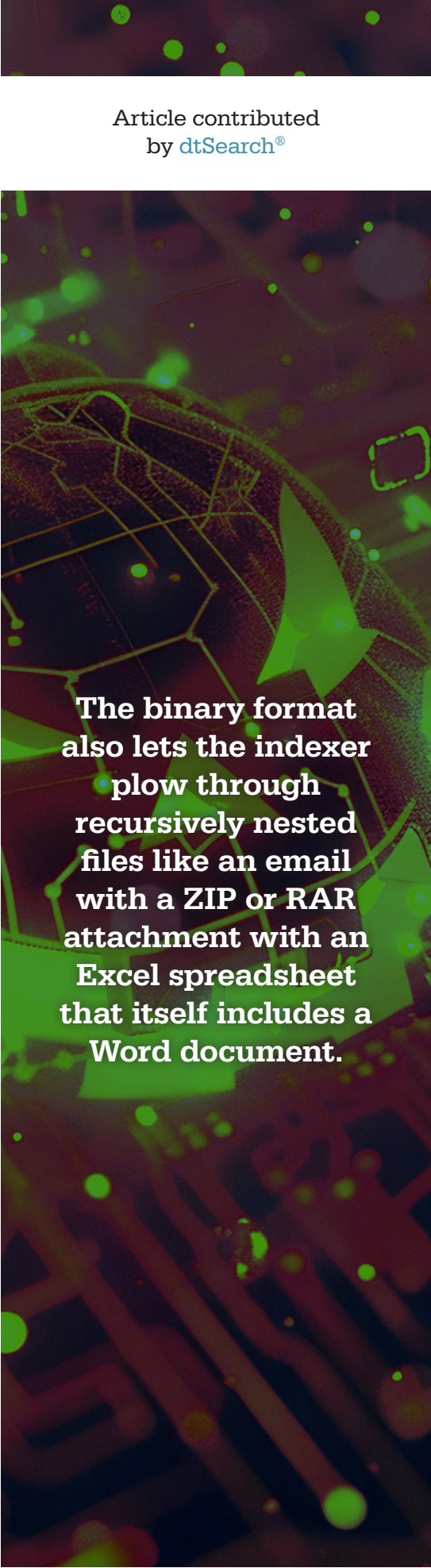
And are there other benefits to binary format access?

While text that blends in with its background color like midnight black writing against a midnight black background is very hard to spot while viewing a file in its native application, such text is as readily apparent as any other text in the binary format. Using the binary format, the indexer can flag "image only" PDFs where the search engine only has the filename and the metadata to work with, but no full-text content. After this identification, simply run the "image only" PDFs through an OCR program like Adobe Acrobat Reader to make them full-text searchable and return them to the indexer.

What about index capacity?

A dtSearch index can hold up to a terabyte of text from mixed local and cloud-based data sources, with no limit on the number of indexes the program can create and instantly search in a concurrent-search environment. And updating an index to reflect new content can proceed without impacting continuing concurrent searching, making it easy to keep indexes up-to-date.

Article contributed
by [dtSearch®](#)



The binary format also lets the indexer plow through recursively nested files like an email with a ZIP or RAR attachment with an Excel spreadsheet that itself includes a Word document.

What types of search features are there?

dtSearch has over 25 different full-text and metadata search options. Enter *ghouls goblins ghosts* as an “any words” natural language search to find items mentioning even one of these terms. Enter *ghouls goblins ghosts* as an “all words” natural language search to locate only items that contain all three of these terms. Boolean and proximity searching can hone in on the phrase *dark night ghouls* or the phrase *Halloween ghosts* within 13 words of *Halloween goblins*. For further refinement, add on a metadata element like subject metadata contains *scary vampire* and not *trick-or-treat*.

What are some other popular search options?

Concept searching extends a search for *ghost* to *apparition*. Fuzzy searching adjusts from 1 to 10 to sift through typographical deviations like *Halloween* mis-typed or mis-OCRed as *Hallomeen*. dtSearch can also search for numbers or numeric ranges as well as dates or date ranges, including automatically extending across common date variants like *October 31, 2024*, *Oct 31, 2024* and *10/31/24*. dtSearch can even identify any credit numbers lurking in data. And dtSearch works with Unicode to support hundreds of international languages. A file can cycle through European languages, right-to-left languages like Hebrew or Arabic, double-byte Asian text, then back to European languages, and dtSearch and Unicode will track the whole procession.

And relevancy-ranking?

By default, dtSearch applies vector-spaced relevancy-ranking. In an “any words” search for *ghouls goblins ghosts*, if *goblins* and *ghosts* are prevalent across indexed data but *ghouls* rare, items with *ghouls* will get a higher relevancy rank, with the densest-mentioning files coming out on top. Or override default ranking through custom positive and negative variable term weighting giving *goblins* a positive weight of 4, *ghosts* a negative weight of 2, and *ghouls* a positive weight of 7 but only for appearances in specific metadata or

Article contributed
by dtSearch®




dtSearch can even identify any credit numbers lurking in data. And dtSearch works with Unicode to support hundreds of international languages. A file can cycle through European languages, right-to-left languages like Hebrew or Arabic, double-byte Asian text, then back to European languages, and dtSearch and Unicode will track the whole procession.

near the top or bottom of a file. Or at anytime, instantly re-sort search results by some unrelated metric like file date or file location. Whatever the sorting, dtSearch displays the full-text of files with highlighted hits for easy navigation as if by the light of a full Halloween moon.

Final thoughts?

Let enterprise search proactively locate your enterprise data ghouls ... before they reach out from the beyond to find you. Download fully-functional 30-day evaluation versions at dtSearch.com.

Article contributed
by [dtSearch®](http://dtSearch.com)



**Let enterprise search
proactively locate
your enterprise data
ghouls ... before they
reach out from the
beyond to find you.
Download fully-
functional 30-day
evaluation versions
at dtSearch.com.**